NEW YORK UNIVERSITY
INSTITUTE OF MATHEMATICAL SCIENCES
25 W. [illegible], New York 3, N. Y.

AEC Computing and Applied Mathematics Center

NYO-8673

AEC RESEARCH AND DEVELOPMENT REPORT

TID - 4500
13th Ed.

NYO - 8673
PHYSICS AND
MATHEMATICS

RELAXATION METHODS FOR LINEAR EQUATIONS

by

Samuel Schechter

September 1, 1958

Institute of Mathematical Sciences

NEW YORK UNIVERSITY
NEW YORK, NEW YORK

NYO-8673
c.1

This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

A. Makes any warranty or representation, express or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission to the extent that such employee or contractor prepares, handles or distributes, or provides access to, any information pursuant to his employment or contract with the Commission.

AEC Computing and Applied Mathematics Center
Institute of Mathematical Sciences
New York University

TID - 4500
13th Ed.

NYO - 8673
PHYSICS AND
MATHEMATICS

RELAXATION METHODS FOR LINEAR EQUATIONS

by

Samuel Schechter

September 1, 1958

Contract No. AT(30-1)-1480

NEW YORK UNIVERSITY
INSTITUTE OF MATHEMATICAL SCIENCES
LIBRARY
25 Waverly Place, New York 3, N. Y.

NYO - 8673
PHYSICS AND
MATHEMATICS

ABSTRACT

Methods of group relaxation analogous to those of Gauss and Southwell are analized and shown to converge for symmetric positive definite matrices. The methods treated allow for a repartitioning and change in norm with each relaxation step. A method of approximate relaxation is introduced which includes over- or underrelaxation and an application is made to compound iteration processes. Explicit error estimates are given in certain cases and examples of special matrices are treated. This extends the theorems of Temple and Ostrowski to group relaxation methods and some of these are seen to be feasible for use by automatic digital computers.

Relaxation processes with arbitrary orderings are shown to converge for positive definite symmetric matrices and also for a special class of matrices with non-negative inverses.

New York University
INSTITUTE OF MATHEMATICAL SCIENCES
New York 3, N. Y.

TABLE OF CONTENTS

Page

Abstract 2

Section

1. Introduction 4
2. Notation and Definitions 6
3. Main Theorem 11
4. Remarks 17
5. Application to Compound Iterations 21
6. Arbitrary Orderings 26
7. Convergence Rate 36
8. Examples 38

Bibliography 45

NEW YORK UNIVERSITY
INSTITUTE OF MATHEMATICAL SCIENCES
LIBRARY
25 Waverly Place, New York 3, N. Y.

RELAXATION METHODS FOR LINEAR EQUATIONS

1. Introduction

Let $A = (a_{ij})$ $i,j = 1,2,\ldots,n$ be a real non-singular matrix of order $n > 1$ and consider the problem of solving the equation

$$r(u) \equiv Au - b = o \tag{1.1}$$

for the column vector u, given b, by an iterative procedure. Such a procedure was introduced by Gauss [4], to solve (1.1) and is commonly referred to as a relaxation method [3], [11]. Among the matrices to which the method is applicable are those obtained from the discretization of certain boundary value problems. Southwell [9], who gave the method its name, and his collaborators have applied relaxation methods extensively and successfully to many such problems.

The method, as described by Gauss, assumes an arbitrary initial guess $u^{(o)}$ is given. One of the n scalar equations of (1.1), say the i^{th}, is selected and the i^{th} component $u_i^{(o)}$ is replaced by the number $u_i^{(1)}$ such that the i^{th} equation is satisfied. The equation selected is usually one which is "worst," measured by some fixed norm of comparison. This procedure is then repeated indefinitely.

A method where the choice of i is made according to the largest of the $|r_i^{(o)}|$, where $r^{(o)} = Au^{(o)} - b$ is called the residual vector, was shown to converge to a solution, by Temple [11], for positive definite symmetric matrices. This method was extended by Ostrowski [6] to allow for incomplete relaxation, that is under- or over-relaxation, and a choice of three norms at each step. He also gives an error estimate for this process.

The theorems given below extend these methods to group (or block) relaxation, that is, where a subset of components of $u^{(o)}$ are modified at each step by the inversion of a "small" matrix. The incomplete relaxation is then replaced by, what may be called, approximate relaxation.

Group relaxation techniques are found to be more feasible to code for an automatic computer than the scalar processes and seem to be competetive with cyclic processes in time of operation.

A proof is also given for the convergence of a process where the groups are chosen, at each step, in an arbitrary manner. The class of matrices to which this applies includes the discrete Laplacian and is related to a class treated by Ostrowski [7]. A free steering theorem is also given for positive definite symmetric matrices.

2. Notation and Definitions

Let Z denote the set of integers $(1,2,\ldots,n)$ and denote by S the class of non-empty proper subsets g of Z such that if $g = (i_1,i_2,\ldots,i_\nu)$ then $i \le i_1 < i_2 < < \ldots < i_\nu \le n$. For each pair $g,h \in S$, $h = (j_1,\ldots,j_\mu)$ we denote a submatrix of A by

$$A_{gh} = \begin{bmatrix} a_{i_1 j_1} & a_{i_1 j_2} & \cdots & a_{i_1 j_\mu} \\ \vdots & & & \\ a_{i_\nu j_1} & a_{i_\nu j_2} & \cdots & a_{i_\nu j_\mu} \end{bmatrix}$$

The matrix A_{gg} is called a principal submatrix of A for any $g \in S$.

Let E_n denote the linear n-dimensional space of column vectors $u = (u_1,u_2,\ldots,u_n)^T$ then for a given $g \in S$ $u_g = (u_{i_1},u_{i_2},\ldots,u_{i_\nu})^T$ will denote a ν-dimensional subvector of u.

Let $\pi = (g_1,g_2,\ldots,g_N)$ denote a set of distince $g_i \in S$, $1 \le i \le N$ such that

$\bigcup_{k=1}^{N} g_k = Z$ then π will be called a covering of Z (or simply a covering). If π is a covering of Z then π is called a partition of Z (or simply a partition) if $g_i \cap g_j = 0$, $i \ne j$, $1 \le i,j \le N$, and

$\bigcup_{k=1}^{N} g_k = Z$.

If π is any partition then (1.1) may be written as

$$r_{g_i}(u) \equiv \sum_{k=1}^{N} A_{g_i g_k} u_{g_k} - b_{g_i} = 0$$

$$g_i \in \pi , \qquad i = 1,2,\ldots,N .$$

In particular, if we denote the complement of g by $g' = \mathbb{Z}-g$ then for the partition $\pi = (g,g')$, (1.1) becomes, for any $g \in \mathbb{S}$

$$\begin{aligned} r_g(u) &\equiv A_{gg}u_g + A_{gg'}u_{g'} - b_g = 0 \\ r_{g'}(u) &\equiv A_{g'g}u_g + A_{g'g'}u_{g'} - b_{g'} = 0 \end{aligned} \tag{2.1}$$

Let $K = (c_{\alpha\beta})$, $\alpha = 1,2,\ldots,\nu$; $\beta = 1,2,\ldots,\mu$; $1 \le \mu,\nu \le n$ be a given matrix then for the g and h given above we define the <u>inflated matrix of K</u> to be

$$[K]_{gh} = (c_{pm}) \qquad p,m = 1,2,\ldots,n \text{ where}$$

$$c_{pm} = \begin{cases} c_{\alpha\beta} & \text{if } p \in g \text{ and } m \in h \text{ and } p = i_\alpha,\ m = j_\beta \\ 0 & \text{otherwise .} \end{cases} \tag{2.2}$$

If $\mu = 1$, i.e., K is a column vector, then write $[K]_g$.

Let K be a square matrix of the same order as A_{gg} then it follows readily from the above definitions that for any $r \in E_n$

$$r^T[K]_{gg}\, r = r_g^T\, K\, r_g \tag{2.3}$$

$$[K]^T_{gg}\, A[K]_{gg} = [K^T A_{gg} K]_{gg} \tag{2.4}$$

for any $g \in S$.

Let $\{g(p)\}_{p=0}^{\infty}$ be a sequence of sets $g(p) \in S$. To each $p = 0,1,2,\ldots$ associate the principal submatrix $A_{g(p)g(p)}$, which we assume to be non-singular, and let $B_p = A^{-1}_{g(p)g(p)}$.

We will often also write $\{x_p\}$ for the infinite sequence $\{x_p\}_{p=0}^{\infty}$, with p understood to range from 0 to ∞.

Let $\{C_p\}$ be a sequence of matrices of the same orders as the $\{B_p\}$ respectively, and let

$$K_p = B_p + C_p . \tag{2.5}$$

For any column vector $u^{(o)} \in E_n$ we define an <u>approximate relaxation process with ordering</u> $g(p)$ by

$$\begin{aligned} u_g^{(p+1)} &= u_g^{(p)} - K_p r_g^{(p)} \qquad g = g(p) \\ u_{g'}^{(p+1)} &= u_{g'}^{(p)} \end{aligned} \tag{2.6}$$

where $r_g^{(p)} = r_g(u^{(p)})$.

We denote the solution of (1.1) by u^* and we shall examine the question of the convergence of $\{u^{(p)}\}$ to u^*.

If $\{u^{(p)}\}$ converges to u^* for every initial guess $u^{(o)}$ and every b we will say that the process (2.6) is <u>completely convergent</u>.

The process of applying (2.6) will be called a relaxation step or briefly a relaxation. The particular relaxations obtained when

$$C_p = \varepsilon_p B_p \qquad p = 0,1,2,\ldots , \tag{2.7}$$

where ε_p is a given sequence of scalars, is referred to by Ostrowski [6] as incomplete relaxation i.e. overrelaxation ($0 < \varepsilon_p$) or underrelaxation ($\varepsilon_p < 0$). (The numbers $q_p = {}_p = 1 + \varepsilon_p$ are sometimes called relaxation factors).

For the general process defined by (2.6) we obtain from (2.2) that the residuals are iterated by the rule:

$$r_h^{(p+1)} = r_h^{(p)} - A_{hg} K_p r_g^{(p)} , \quad h = g, g' , \quad g = g(p) \tag{2.8}$$

For the case of incomplete relaxation we get from (2.7) that

$$r_g^{(p+1)} = - \varepsilon_p \, r_g^{(p)}$$

and for "exact" relaxation, that is $\varepsilon_p = 0$, $r_g^{(p+1)} = 0$.

If C_p is given by (2.7) and $\pi = (g_1, g_2, \ldots, g_N)$ is a given partition then the process given by (2.6) is the well known (see for instance [5], [6], [14], [16] for group iterations and [15], [17], for the scalar case) cyclic linear group iteration if $g(p)$ is defined by

$$\begin{aligned} g(p) &= g_{p+1} && 0 \le p < N \\ g(p) &= g(p-N) && N \le p \quad . \end{aligned} \tag{2.9}$$

(Most treatments of this assume ε_p to be a fixed number.)

The case to be treated here is one where $g(p)$ is computed by a comparison of components of the residual vector.

Let $\pi = (g_1,\ldots,g_N)$ be a given covering and $\|*\|$ a given norm for E_n then for any $r \in E_n$ define

$$\bar{g} = \zeta(\pi, r, \|*\|) \tag{2.10}$$

as the set $g_i \in \pi$ (with smallest subscript i) such that

$$\|[r_{\bar{g}}]_{\bar{g}}\| = \max_{1\le k\le N} \|[r_{g_k}]_{g_k}\| \ . \tag{2.11}$$

Let $\{\pi_p\}$ be a given sequence of coverings such that

$$\pi_p = (g_1^{(p)},\ldots,g_{N_p}^{(p)}) \ , \qquad 1 < N_p \le N \ . \tag{2.12}$$

Let $\{\|*\|_{(p)}\}$ be a given sequence of norms and $\{\eta_p\}$, $\{\tau_p\}$ sequences of positive constants such that for any $r \in E_n$

$$\begin{aligned} \eta_p \tau_p |r|^2 \le \|r\|^2_{(p)} \le \eta_p |r|^2 \\ 0 < \eta_p \tau_p \le \eta_p \qquad p = 0,1,2,\ldots \end{aligned} \tag{2.13}$$

where $|r|^2 = r^T r$. Let the ordering $\bar{g}(p)$ be defined by

$$g = \bar{g}(p) = \zeta(\pi_p, r^{(p)}, \|*\|_{(p)}). \tag{2.14}$$

Under these conditions we say that the relaxation process (2.6) is residually ordered, or briefly a residually ordered process (r.o.p.).

3. Main Theorem

Unless specified otherwise A will henceforth be assumed to be symmetric, although some of the results hold also for Hermitian matrices. For the r.o.p. defined above we prove:

Theorem 1. *For the ordering* $g = g(p)$ *defined by the above r.o.p. let the matrices* A_{gg}, $g \in \pi_p$, $p = 0,1,2,\ldots$ *be positive definite. Let a sequence of matrix valued functions* $C_p(r^p, A, g) = C_p$, *and non-negative numbers* δ_p, $p = 0,1,2,\ldots$ *exist, such that the* C_p *are of the same orders as the* A_{gg} *and such that they satisfy the conditions:*

$$(3.1) \qquad (C_p r_g^{(p)})^T A_{gg} C_p r_g^{(p)} \leq \delta_p \, (r_g^{(p)})^T B_p r_g^{(p)} , \quad g = g(p)$$

$$(3.2) \qquad 0 \leq \delta_p < 1 \qquad \text{for} \quad p = 0,1,2,\ldots$$

$$(3.3) \qquad \sum_{p=0}^{\infty} (1-\delta_p)\tau_p \quad \text{diverges.}$$

The r.o.p. defined by (2.6) and (2.14) is then completely convergent if and only if A *is positive definite.*

If (3.3) is replaced by the assumption that there exist positive numbers δ *and* τ *such that*

$$(3.4) \qquad 0 \leq \delta_p \leq \delta < 1$$

$$(3.5) \qquad 0 < \tau \leq \tau_p \leq 1$$

then the above conclusion is still valid and the sequence $\{u^p\}$ *converges to* u^* *like a geometric series, i.e., there*

exist positive constants θ and α such that

$$|u^{(p)} - u^*|^2 < \theta\, \alpha^p\, |u^{(o)} - u^*|^2$$

(3.6)

$$0 < \alpha < 1 \quad .$$

The proof of this theorem depends on several lemmas:

Lemma 1. Let L_p be a sequence of matrices of order n, then for an arbitrary $e^{(o)} \varepsilon E_n$ define the sequence $e^{(p)}$ by

$$e^{(p+1)} = L_p e^{(p)} , \qquad p = 0,1,2,\ldots . \tag{3.7}$$

Let $V(e) = e^T A e$ for $e \varepsilon E_n$ and assume that

$$V(e^{(p+1)}) \leq V(e^{(p)}) , \qquad p = 0,1,2,\ldots . \tag{3.8}$$

If $\lim_{p\to\infty} e^{(p)} = 0$ for every initial $e^{(o)}$ then A is positive definite.

Proof: Assume that for some $e^{(o)} \neq 0$, $V(e^{(o)}) < 0$ then for the sequence defined by (3.7)

$$0 > V(e^{(o)}) \geq V(e^{(1)}) \geq \ldots \geq V(e^{(p)}) \geq \ldots$$

Since $e^{(p)} \longrightarrow 0$ and since A is non-singular $V(e^{(p)}) \longrightarrow 0$ which is impossible. We therefore have that $V(e) \geq 0$ for all $e \varepsilon E_n$. Since A is symmetric and non-singular $V(e) \neq 0$ for $e \neq 0$ and the lemma is proved.

Lemma 2. Let A be positive definite and, for a given $e^{(o)} \varepsilon E_n$, let the L_p and $e^{(p)}$ be defined as in Lemma 1

by (3.7). Assume that there exists a sequence $\{\alpha_p\}$ such that for any $e^{(o)}$

$$0 \leq \alpha_p < 1 \tag{3.9}$$

$$V(e^{(p+1)}) \leq \alpha_p V(e^{(p)}) \; ; \qquad p = 0,1,2,\ldots \tag{3.10}$$

and assume that

$$\sum_{p=0}^{\infty} (1-\alpha_p) \quad \text{diverges} \; , \tag{3.11}$$

then $\lim_{p\to\infty} e^{(p)} = 0$ for any $e^{(o)}$.

If instead of (3.11) it is assumed that

$$0 \leq \alpha_p \leq \alpha < 1 \tag{3.12}$$

then $e^{(p)} \longrightarrow 0$ like a geometric series.

Proof: Since A is positive definite, there exist positive numbers λ, Λ such that

$$\lambda \cdot e^T e \leq V(e) \leq \Lambda \cdot e^T e \quad .$$

From this and (3.10) it follows that

$$\lambda |e^{(p+1)}|^2 \leq V(e^{(p+1)}) \leq \alpha_p V(e^{(p)}) \leq \beta_p |e^{(o)}|^2 \Lambda$$

where we set $\beta_p = \prod_{k=0}^{p} \alpha_k$. From (3.9) we obtain that $\alpha_p < \exp(\alpha_p - 1)$, so that from (3.11) we get that $\beta_p \longrightarrow 0$, and the convergence is proved.

If on the other hand (3.12) is true then

(3.13) $$|e^{(p)}|^2 < a^p \frac{\Lambda}{\lambda} |e^{(o)}|^2 ,$$

or the convergence is geometric and the lemma is proved.

Proof of Theorem 1: Let $u^{(o)} \varepsilon E_n$ be arbitrary and define first the process

(3.14) $$u^{(p+1)} = u^{(p)} - Q_p r^{(p)}$$

where $Q_p = Q_p(u^{(o)},u^{(1)},\ldots,u^{(p)},A,b)$, $p = 0,1,\ldots$ is a given sequence of matrices of order n. Let u^* be the solution of (1.1) and denote the p^{th} error by $e^{(p)} = u^{(p)}-u^*$. Let $L_p^* = I - Q_pA$, where I is the unit matrix of order n, and assume that

(3.15) $$Q_p(u^{(o)},\ldots,u^{(p)},A,b) = Q_p(e^{(o)},\ldots,e^{(p)},A,0) ,$$

$$p = 0,1,2,\ldots$$

for any $u^{(o)}$ and b . It then follows from (3.15) and from $u^{(p+1)} - u^* = L_p^* e^{(p)}$ that the convergence of $u^{(p)}$ to u^* is equivalent to the convergence of $e^{(p)}$ to 0 where $e^{(p)}$ is given by (3.7) and

$$L_p = L_p^*(e^{(o)} + u^*, \ldots, e^{(p)} + u^*,A,b), \text{ that is}$$

(3.16) $$e^{(p+1)} = (I - Q_pA)\, e^{(p)} ,$$

(3.17) $$r^{(p+1)} = (I - AQ_p)\, r^{(p)} .$$

If we now set

(3.18) $$Q_p = [K_p]_{gg} \qquad g = g(p) \ ,$$

where K_p is given by (2.5), then it is clear from (2.2) that the relaxation processes (2.6) are of the form (3.14). Since K_p depends on $u^{(p)}$ and b only through $r^{(p)}$ in the r.o.p., (3.15) is satisfied. It is therefore sufficient to show that the conditions of Lemmas 1 and 2 are satisfied for the L_p given above.

To show the necessity part of Theorem 1 we verify (3.8). We will find it convenient at times to drop the superscript p and write $()'$ for $()^{(p+1)}$, g for $\bar{g}(p)$.

We note that

(3.19) $$V(e) = e^T Ae = r^T A^{-1} r$$

and for $e' = e - Q_p r = L_p e$

$V(e') = e^T Ae - 2e^T AQ_p r + (Q_p r)^T AQ_p r$ so that, for Q_p and K_p given by (3.18) and (2.5)

$$\begin{aligned} V(e') - V(e) &= -2r^T [K_p]_{gg} r + r^T [K_p]_{gg}^T \, A [K_p]_{gg} r \\ &= -2r_g^T K_p r_g + r^T [K_p^T A_{gg} K_p]_{gg} r \\ &= r_g^T (K_p^T A_{gg} K_p - 2K_p) \, r_g \\ &= r_g^T (C_p^T - C_p - B_p + C_p^T A_{gg} C_p) \, r_g \\ &= r_g^T (C_p^T A_{gg} C_p - B_p) \, r_g \end{aligned}$$

from (2.3) and (2.4).

Since the B_p are positive definite we get from (3.1), (3.2) that

$$V(e') - V(e) \leq (\delta_p - 1)\, r_g^T B_p r_g \leq 0 \,. \tag{3.20}$$

(This of course is valid for any ordering $g = g(p)$ such that (3.1) and (3.2) hold). Thus we have the conditions of Lemma 1 satisfied and A is positive definite.

Assume now that A is positive definite and denote, for any $g \in S$, $\Lambda(A_{gg}) = \Lambda_g$, $\lambda(A_{gg}) = \lambda_g$ to be the largest and smallest eigenvalue of A_{gg} respectively. In particular set $\Lambda(A) = \Lambda$, $\lambda(A) = \lambda$. It is then easily seen that

$$\lambda \leq \lambda_g \leq \Lambda_g \leq \Lambda \tag{3.21}$$

and A_{gg} is positive definite.

For π_p given by (2.12) write $\pi_p = (g_1, g_2, \ldots, g_{N_p})$ then for $g = \bar{g}(p)$

$$\Lambda_g r_g^T B_p r_g \geq |r_g|^2 = |[r_g]_g|^2 \geq \frac{1}{\gamma_p} \|[r_g]_g\|^2_{(p)} \geq$$

$$\geq \frac{1}{N_p \gamma_p} \sum_{k=1}^{N_p} \|[r_{g_k}]_{g_k}\|_{(p)} \geq \frac{\tau_p}{N_p} |r|^2 \geq \frac{\lambda \tau_p}{N_p} V(e) \,.$$

Thus it follows from (3.20) that

$$V(e^{(p+1)}) \leq \alpha_p V(e^{(p)}) \quad \text{where} \tag{3.22}$$

$$\alpha_p = 1 - (1-\delta_p)\, \tau_p \frac{\lambda}{N_p \Lambda_{\bar{g}(p)}} \,. \tag{3.23}$$

Since the Λ_g and N_p are finite in number we get from (3.3) that $\sum_{p=0}^{\infty} (1-\alpha_p)$ diverges and the convergence follows from Lemma 2.

If on the other hand τ_p and δ_p satisfy (3.4) and (3.5) then let

$$\alpha = 1-(1-\delta)\tau\frac{\lambda}{N\mu}, \quad \mu = \max_p \max_{g\varepsilon\pi_p} \Lambda_g . \tag{3.24}$$

Since $0 \leq \alpha_p \leq \alpha < 1$ we obtain from (3.13) that (3.6) is valid with $\theta = \Lambda/\lambda$, and the theorem is proved.

We will refer to the α_p defined by (3.23) as the reduction factors of the r.o.p.

4. Remark 1

The following are examples of norms $\|*\|_{(p)}$ that may be used in practice:

(1) (4.1) $\qquad \|r\| = |r|, \quad \tau = \eta = 1$

(2) For a given covering $\pi = (g_1, g_2, \ldots, g_N)$ let

$$\mu = \max_{1 \le k \le N} \wedge_{g_k}, \qquad \sigma = \min_{1 \le k \le N} \lambda_{g_k}$$

(4.2) $$\|r\|^2 = \sum_{k=1}^{N} r_{g_k}^T \, A_{g_k g_k}^{-t} \, r_{g_k}, \quad t \ge 1$$

then $\eta = \dfrac{N}{\sigma^t}$, $\quad \tau = \left(\dfrac{\sigma}{\mu}\right)^t \dfrac{1}{N}$.

If π is a partition, then $\eta = \dfrac{1}{o^t}$, $\quad \tau = \left(\dfrac{\sigma}{\mu}\right)^t$.

(3) (4.3) $$\|r\| = \sum_{j=1}^{n} |r_j|, \ \eta = n, \ \tau = \frac{1}{n}.$$

(4) (4.4) $$\|r\| = \max_{1 \le j \le n} |r_j|, \ \eta = 1, \ \tau = \frac{1}{n}.$$

These norms may of course be mixed and used in any order.

Remark 2.

For the particular case where the coverings π_p are partitions and are all the same $\pi_p \equiv \pi = (g_1, g_2, \ldots, g_N)$ we may without loss assume that the g_k are consecutive; that is $g_1 = (1, 2, \ldots, i_1)$, $g_2 = (i_1 + 1, i_1 + 2, \ldots, i_2), \ldots,$ $g_N = (i_{N-1}+1, i_{N-1}+2, \ldots, n)$. It is then convenient to denote

$A_{g_i g_j} = A_{ij}$; $u_{g_i} = u_i$; $\bigwedge_{g_i} = \bigwedge_i$, $\lambda_{g_i} = \lambda_i$ and the ordering $g(p) = g_{i(p)}$ by the integer $i(p)$, $1 \leq i(p) \leq N$.

The matrix A may be written in the form:

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1N} \\ \vdots & & & \\ A_{N1} & A_{N2} & \cdots & A_{NN} \end{bmatrix}$$

The iteration (2.6) for a given ordering $i(p)$ may then be written in the form

$$(4.5) \qquad \begin{aligned} u_j^{(p+1)} &= u_j^{(p)} - K_p r_j^{(p)} && j = i(p) \\ u_j^{(p+1)} &= u_j^{(p)} && j \neq i(p) \end{aligned}$$

where $K_p = A_{ii}^{-1} + C_p$ and the residuals iterate by the rule

$$(4.6) \qquad r_k^{(p+1)} = r_k^{(p)} - A_{ki} K_p r_i^{(p)} , \qquad i = i(p) .$$

Let the ordering $g = \bar{g}(p)$ of the r.o.o.p. be replaced by

$$(4.7) \qquad i = m(p)$$

that is, i is the smallest integer such that $\| [r^{(p)}]_{g_i} \, g_i \|_{(p)}$ is maximal, then this is a r.o.o.p. but with a fixed partition.

For the special case where C_p is given by (2.7), (the incomplete relaxation case), such that $-1 < \varepsilon_p < 1$,

$p = 0,1,2,\ldots$; let $\delta_p = \varepsilon_p^2$. It then follows immediately that (3.1), (3.2) are satisfied and if

$$\sum_{p=0}^{\infty} (1-\varepsilon_p^2)\tau_p$$

diverges, all the conditions of Theorem 1 are satisfied.

Ostrowski [6] has proved the convergence of this r.o.p. for Hermitian matrices, under the assumption that $N = n$, that is A_{ij} is simply the scalar a_{ij} and where the norms are restricted to be any mixture of the three norms (1) and (2), $t = 1,2$.

The scalar process obtained with the further restriction that $\varepsilon_p \equiv 0$ or $C_p \equiv 0$ was shown to converge by Temple [11], under the assumption that norm (1) is used exclusively.

Theorem 1 allows <u>any</u> group to be relaxed if it contains an equation with maximal residual in absolute value. This is obtained if each π_p consists of all the elements $g \in S$.

It should also be noted from (3.6) that the error bound is dependent on the size of the number Λ/λ. This is what is called the P-condition number of the matrix A (see J. Todd [12]) and may get quite large for poorly conditioned matrices.

<u>Remark 3</u>.

Suppose that between each relaxation step of the r.o.p. of Theorem 1 a finite number of iterations of any form are taken which do not increase $V(e)$. These may for instance

be relaxations of the form (2.6), where the C_p satisfy (3.1) and (3.2), and are taken in any order. This would amount to extending Lemma 2 to require only that $0 \leq a_p \leq 1$ and that only for an infinite subsequence is (3.9) valid. This is easily done and implies that, except for the error estimate (3.6), the conclusions of Theorem 1 remain unchanged.

5. Application to Compound Iterations

It will now be shown how the approximate relaxation method of Theorem 1 may be applied to the practical problem of carrying out the individual relaxation steps. When given by (2.6) these involve the inversion of the matrix A_{gg} or solving a set of linear equations. Even if the inverse of this matrix is known explicitly (which often is not the case) it may be more practical to compute $u^{(p+1)}$ only approximately. Even in the scalar processes it is a common practice, for instance, to compute $u^{(p+1)}$ only to a few significant digits.

Since the matrix K_p may be viewed as an approximation to A_{gg}^{-1}, with the allowable error measured by δ_p a certain amount of error is allowed in the computation of $u^{(p+1)}$. It may in fact be convenient, as is done in practice, to compute $u^{(p+1)}$ itself by some iterative process. These subsidiary iterations are sometimes called inner iterations. It is indicated below how these iterative processes may be compounded.

We consider the following compound iterative method: Let $u^{(o)}$ be arbitrary and let π_p, C_p, δ_p, $\|*\|_{(p)}$, η_p, τ_p, $g(p)$ be defined as in the r.o.p. to satisfy (3.1) and (3.2).

Suppose $u = u^{(p)}$, $r = r^{(p)}$ have been computed, we will assume $u' = u^{(p+1)}$ to be computed as follows: Determine $g = \bar{g}(p)$ as given by (2.14) and set $u'_{g'} = u_{g'}$, $g' = Z-g$. Let v^* be the solution of

$$s(v) \equiv A_{gg}(v-u_g+C_p r_g) + r_g = 0 \tag{5.1}$$

and let $\tilde{v}$ be an approximate solution of (5.1) such that $\tilde{v} - v^* = w = w(p,A,r_g)$ depends at most on p,A and r_g. We then set $u_g^1 = \tilde{v}$ and the process is continued.

Theorem 2. Let $A_{gg}, g = \bar{g}(p)$ be positive definite and for the compound process defined above, assume that there exists positive numbers γ_p, $p = 0,1,2,\ldots$ such that

$$0 \leq \delta_p < \gamma_p < 1 \qquad p = 0,1,2,\ldots \tag{5.2}$$

$$|w| \leq \frac{\gamma_p^{1/2} - \delta_p^{1/2}}{\Lambda_g} |r_g| \qquad g = \bar{g}(p) \tag{5.3}$$

and such that

$$\sum_{p=0}^{\infty} (1-\gamma_p)\tau_p \quad \text{diverges} \tag{5.4}$$

then the above compound process is completely convergent if and only if A is positive definite.

If instead of (5.4) the τ_p satisfy (3.5) and

$$0 < \gamma_p \leq \gamma < 1 \qquad p = 0,1,2,\ldots \tag{5.5}$$

then $u^{(p)}$ converges to u^* like a geometric series.

Proof: Let $\theta = \frac{\gamma_p^{1/2} - \delta_p^{1/2}}{\Lambda_g}$ then we exhibit a set of matrices $\tilde{C}_p = \tilde{C}_p(r^{(p)},A,g)$ and numbers $\tilde{\delta}_p$ such that the conditions of Theorem 1 are satisfied.

Let D_p be a matrix such that $D_p r_g = w$. This matrix exists since $r_g \neq 0$ for $g = \bar{g}(p)$, and depends at most on

r, A and g. We may therefore write $\tilde{v} = v^{*} + D_p r_g$ or $u'_g = u_g - (B_p + C_p - D_p)r_g$, $g = \bar{g}(p)$. Let $\tilde{C}_p = C_p - D_p$ and $d = r_g^T B_p r_g$. From (3.1), (5.3) and the consequent inequalities $|w|^2 \leq \theta^2 \lambda_g d$, $|A_{gg} C_p r_g|^2 \leq \Lambda_g \delta_p d$ we obtain that

$$(C_p r_g)^T A_{gg} (\tilde{C}_p r_g) = (C_p r_g)^T A_{gg} (C_p r_g) - 2w^T A_{gg} C_p r_g + w^T A_{gg} w$$

$$\leq \delta_p d + 2|w|\,|A_{gg} C_p r_g| + \Lambda_g |w|^2$$

$$\leq (\delta_p + 2\theta \Lambda_g \sqrt{\delta_p} + \lambda_g^2 \theta^2) d = (\sqrt{\delta_p} + \lambda_g \theta)^2 d .$$

If we let $\tilde{\delta}_p = (\sqrt{\delta_p} + \lambda_g \theta)^2$ then $\tilde{\delta}_p = \gamma_p$, and, since γ_p satisfies (5.2) and (5.4) the conditions of Theorem 1 are satisfied. If γ_p satisfies (5.5) then clearly (3.4) is satisfied and the theorem is proved.

For the particular case where $\tilde{v}$ is obtained by successive relaxations we have

Corollary 1. Let a subsidiary r.o.p., which satisfies the conditions of Theorem 1 be given for (5.1) such that the initial guess is $v^{(o)} = u_g^{(p)} - C_p r_g^{(p)}$. Let $q = q_p \geq 1$ be the smallest integer such that

$$\tilde{\beta}_{q-1} = \prod_{t=0}^{q-1} \tilde{\alpha}_t \leq \left(\frac{\lambda_g}{\Lambda_g}\right)^2 \cdot (\sqrt{\gamma_p} - \sqrt{\delta_p})^2 \tag{5.6}$$

where the $\tilde{\alpha}_t$ are the reduction factors of the subsidiary r.o.p. and the γ_p satisfy (5.2) and (5.4). If we set $\bar{v} = v^{(q)}$ in the above compound process then it is completely convergent if and only if A is positive definite.

Proof: We note that $\tilde{v} - v^{*} = w$ is the error after q

inner iterations and is dependent at most on A, g and on the initial error $w^{(o)} = B_p s(v^{(o)}) = B_p r_g$. To show that (5.3) is satisfied we use (3.19), (3.22) and

$$\lambda_g |w|^2 \le w^T A_{gg} w \le \tilde{\beta}_{q-1} (w^{(o)})^T A_{gg} w^{(o)} = \tilde{\beta}_{q-1} r_g^T B_p r_g$$
$$\le \tilde{\beta}_{q-1} \frac{|r_g|^2}{\lambda_g}$$

to obtain

$$|w| \le \frac{\sqrt{\tilde{\beta}_{q-1}}}{\lambda_g} |r_g| \le \frac{\sqrt{\gamma_p} - \sqrt{\delta_p}}{\Lambda_g} |r_g| \quad .$$

Thus the conditions of Theorem 2 are satisfied and the corollary is proved.

We note that the right hand side of (5.6) depends on the P-condition of A_{gg} and again limits the liberties we may take in the process.

In some cases an explicit approximation for the inverse of a matrix A may be used, for instance $A^{-1} = B \simeq 2I - A = K$. The error is $C = -R^2 B$ where $R = I - A$ and $K = B + C$. This approximation may be used under certain conditions as indicated by the following:

Corollary 2. *Let a r.o.o.p. be defined for (1.1) with the following properties:*

1. *There exist non-negative numbers* ζ_p, $p = 0,1,2,\ldots$ *such that for* $g = \bar{g}(p)$

$$0 < 1 - \zeta_p \le \lambda_g \le \Lambda_g \le \zeta_p + 1 \ , \quad p = 0,1,2,\ldots \tag{5.7}$$

$$\sum_{p=0}^{\infty} (1-\zeta_p^4)\tau_p \quad \text{diverges} \ . \tag{5.8}$$

2. Let $C_p = -R_p^2 B_p$ where $R_p = (I-A)_{gg}$. Then the resulting r.o.p. is completely convergent if and only if A is positive definite.

Proof: Since the C_p clearly depend at most on A and g it remains only to exhibit δ_p, $p = 0,1,2,\ldots$ such that (3.1) and (3.2) are satisfied.

Using the same notation as in Theorem 2 we get

$$(C_p r_g)^T A_{gg} (C_p r_g) = (R_p^4 r_g)^T B_p r_g \leq \bigwedge(R_p^4) d \leq \zeta_p^4 d$$

if we let $\delta_p = \zeta_p^4$ we have form (5.7) and (5.8) that (3.1), (3.2) and (3.3) are satisfied and the corollary is proved.

Similar results hold for higher degree approximations of this type.

6. Arbitrary Orderings

For certain classes of matrices the ordering function $g(p)$ may be chosen in almost any manner and still the resulting relaxation process converges completely. Such matrices have been treated by Ostrowski [7] and he shows them to yield complete convergence for almost all arbitrarily chosen orderings. This he refers to as a free steering method in distinction to the steering obtained in a r.o.p.

We will prove a "free steering" theorem for a class of matrices somewhat similar to Ostrowski's, but allowing approximate relaxations.

We will now drop the assumption that A is symmetric but still assume that A is non-singular. For any m by n matrix $A = (a_{ij})$ we write $A \geq 0$ if $a_{ij} \geq 0$, $1 \leq i \leq m$, $1 \leq j \leq n$; and call A non-negative.

We define (with Collatz [2]) a matrix to be monotone if for every $u \in E_n$ such that $Au \geq 0$ implies that $u \geq 0$. It is clear that A is monotone if and only if $A^{-1} \geq 0$.

Theorem 3. Let A be monotone and $g = g(p)$ any given ordering. Let $\{C_p\}$ be a given sequence of matrices such that for $p = 0,1,2,\ldots$ and $g = g(p)$

i) C_p is of the same order as A_{gg} and $-A_{gg}C_p \geq 0$,

ii) A_{gg} and $K_p = A_{gg}^{-1} + C_p$ are non-singular and $K_p \geq 0$,

iii) $-A_{g'g} \geq 0$.

Assume further that the sequence $\{g(p)\}$ covers Z infinitely often, that is, for each i, $1 \leq i \leq n$, $i \in g(p)$ for infinitely many p.

Then the iterative process (2.6) is completely convergent. The proof depends on the following:

Lemma 3. Let A be monotone and $\{Q_p\}$ a given sequence of matrices of order n such that

$$Q_p \geq 0 \tag{6.1}$$

$$T_p = I - AQ_p \geq 0 \qquad p = 0,1,2,\ldots . \tag{6.2}$$

Then the $\{u^{(p)}\}$ defined by (3.14) converges to some vector in E_n for any $u^{(o)}$ and any b.

Proof: We may as before assume $b = 0$. Assume first that $e^{(o)}$ is given such that

$$r^{(o)} = Ae^{(o)} \geq 0 \ . \tag{6.3}$$

Since A is monotone $e^{(o)} \geq 0$ and from (3.17) and (6.2) it follows that $r^{(p)} \geq 0$ and $e^{(p)} \geq 0$ for $p = 0,1,2,\ldots$. Since $e^{(p+1)} - e^{(p)} = -Q_p r^{(p)} \leq 0$ we have that $0 \leq e^{(p+1)} \leq e^{(p)} \leq \ldots \leq e^{(o)}$ and therefore that $e^{(p)}$ converges to some vector $e^{(\infty)}$.

For the general case where $e^{(o)}$ is arbitrary, let $Ae^{(o)} = \tilde{v} - \tilde{w}$ such that $\tilde{v} \geq 0$ and $\tilde{w} \geq 0$. Let $Av^{(o)} = \tilde{v}$ and $Aw^{(o)} = \tilde{w}$ then $v^{(o)} \geq 0$ and $w^{(o)} \geq 0$. Define the sequences $\{v^{(p)}\}$, $\{w^{(p)}\}$ by $v^{(p+1)} = L_p v^{(p)}$, $w^{(p+1)} = L_p w^{(p)}$

then by the above result $v^{(p)}$, $w^{(p)}$ converge to $v^{(\infty)}$, $w^{(\infty)}$ respectively. Since $e^{(p)} = v^{(p)} - w^{(p)}$ the lemma is proved.

Remark 4.

The same result of Lemma 3 is obtained, by a similar proof, if instead of A monotone it is assumed that $A \geq 0$ and (6.2) is replaced by $I - Q_p A \geq 0$.

Proof of Theorem 2: We note first that $Q_p = [K_p]_{gg} \geq 0$ and since $(Q_p)_{gg'} = (Q_p)_{g'g} = (Q_p)_{g'g'} = 0$ it follows readily that

$$(T_p)_{gg} = I_{gg} - A_{gg}K_p = -A_{gg}C_p \geq 0, \quad (T_p)_{gg'} = 0$$

$$(T_p)_{g'g} = -A_{g'g}K_p \geq 0, \quad (T_p)_{g'g'} = I_{g'g'} \geq 0.$$

Thus the conditions of Lemma 3 are satisfied and $e^{(p)} \longrightarrow e^{(\infty)}$.

To show that $e^{(\infty)} = 0$ we note first that at most a finite number of $g(p)$, $p = 0,1,2,\ldots$ are distinct and some subset of these must occur infinitely often. We denote this (finite) subset by F and for any $r \in E_n$ let $\|r\| = \sum_{k=1}^{n} |r_k|$.

From (3.16) we get that

$$(6.4) \quad e^{(t+k)} - e^{(t)} = \sum_{p=t}^{t+k-1} (e^{(p+1)} - e^{(p)}) = \sum_{p=t}^{t+k-1} (-Q_p r^{(p)}).$$

Assume first that $r^{(o)} \geq 0$, then $r^{(p)} \geq 0$ for $p = o,1,2,\ldots$ Since $e^{(p)}$ converges, it follows from (6.4) that

$$\rho = \|\sum_{p=0}^{\infty} Q_p r^{(p)}\| = \sum_{p=0}^{\infty} \|Q_p r^{(p)}\| = \sum_{p=0}^{\infty} \|K_p r^{(p)}_{g(p)}\|$$

converges. On the other hand, writing K_f for K_p when $g(p) = f$,

$$\rho \geq \sum_{f \in F} \left\{ \sum_{\substack{p=0 \\ p|g(p) = f}}^{\infty} \| K_f r_f^{(p)} \| \right\}$$

and for each $f \in F$ there exists an infinite subsequence $\{p_k\}_{k=0}^{\infty}$ such that $\| K_f r_f^{(p_k)} \| \longrightarrow 0$ as $k \longrightarrow \infty$. Since K_f is non-singular $r_f^{(p_k)} \longrightarrow 0$ for every $f \in F$.

Since each i, $1 \leq i \leq n$, occurs infinitely often there exists an $f \in F$ such that $i \in f$ or that $r_i^{(p_k)} \longrightarrow 0$ as $k \longrightarrow \infty$. Since $r^{(p)} = Ae^{(p)} \longrightarrow Ae^{(\infty)} = r^{(\infty)}$, it follows that $r_i^{(p)} \longrightarrow r_i^{(\infty)} = 0$ as $p \longrightarrow \infty$ or $e^{(\infty)} = 0$.

If $r^{(o)}$ is not non-negative we define $v^{(p)}$, $w^{(p)}$ as in Lemma 3. These, by the preceding argument, converge to zero and the theorem is proved.

Remark 5.

It follows from the proof that the condition of non-singularity on K_f could be replaced by the condition that for each $f \in F$, K_f has no zero column.

It has been shown by de Rham [8] that if a real non-singular matrix A with non-positive off-diagonal elements (i.e., $a_{ij} \leq 0$ $i \neq j$ $1 \leq i,j \leq n$) has no real negative eigenvalues then A is monotone. This is an extension of a theorem of Stieltjes [10] who assumed that A is symmetric and positive definite.

Other sufficient conditions for monotonicity have been given by Collatz [2] and Ostrowski [7].

Birkhoff and Varga [1] have shown that if A is symmetric and $a_{ij} \leq 0$ for $i \neq j$ then A^{-1} exists and is non-negative if and only if A is positive definite.

We therefore have the following

Corollary 3. Let A be a symmetric matrix such that $a_{ii} > 0$, $a_{ij} \leq 0$, $i \neq j$, $1 \leq i,j \leq n$ then the following three statements are equivalent:

(1.) Let $g(p)$ be any ordering that covers Z infinitely often, let C_p be given by (2.7) where $-1 < \varepsilon_p \leq 0$, $p = 0,1,2,\ldots$ then the process (2.6) is completely convergent.

(2.) A is positive definite.

(3.) A is monotone.

Proof: First we note that (1.) implies (2.) by Lemma 1, since the scalar cyclic processes, (that is, where $g(p) = i_p$, $1 \leq i_p \leq n$ and $g(p)$ is given by (2.9) and where $C_p \equiv 0$), satisfies the conditions of that lemma. By de Rham's (or Stieltjes') theorem (2.) implies (3.).

Since, for A monotone, the scalar cyclic processes will be completely convergent by Theorem 2, it follows from Lemma 1 that (3.) implies (2.).

If A is positive definite so are the A_{gg}, $g \in S$, and by de Rham's (or Stieltjes') theorem the A_{gg} are also monotone. Thus all the conditions of Theorem 2 are satisfied

and (2.6) is completely convergent. Thus (2.) implies (1.) and the corollary is proved.

We now prove a free steering theorem for positive definite symmetric matrices.

Let A be a symmetric matrix and $\{g(p)\}$ be any ordering such that it covers Z infinitely often and such that the matrices A_{gg} are positive definite. Let $\{\varepsilon_p\}$ be a sequence of real numbers such that if $\omega_p = 1+\varepsilon_p$ then

$$\gamma \leq \omega_p \leq 2-\gamma \,, \qquad 0 < \gamma < 1 \,. \tag{6.5}$$

Let the matrices $\{C_p\}$ be given by (2.7) then with these assumptions we prove.

Theorem 4. If the ordering $g = g(p)$ is such that, for each p, $r_g^{(p)}$ is not zero then the relaxation process defined by (2.6) and (2.7) is completely convergent if and only if A is positive definite.

Proof: If we let $\delta_p = \varepsilon_p^2$ then (3.1) and (3.2) follow from (2.7). It then follows as in the proof of Theorem 1 that

$$V(e^{(p+1)}) - V(e^{(p)}) = (\varepsilon_p^2-1)(r_g^{(p)})^T B_p r_g^{(p)} \equiv \Delta_p V < 0 \tag{6.6}$$

The necessity then follows from Lemma 1.

To prove the sufficiency we note that the sequence $\{e^{(p)}\}$ is bounded and that if there is only one limit point, that is, if $e^{(p)} \longrightarrow x$ then $x = 0$. This follows from the fact that the $V(e^{(p)})$ decrease monotonically to the limit $V^{\infty} = V(x)$. It then follows from (6.6) that $r_g^{(p)} \longrightarrow 0$. An argument similar to that given in the proof of Theorem 2

then shows that $r^{(p)} \longrightarrow 0$. It therefore suffices to show that $e^{(p)}$ has only one limit point.

Let x be a non-zero limit point of this sequence and W a sufficiently small neighborhood of x which excludes at least one other limit point. It is then clear that for every subneighborhood U of x there exists a consecutive pair $e^{(p)}$ and $e^{(p+1)}$ such that $e^{(p)} \varepsilon U$ and $e^{(p+1)} \notin W$.

Let $y = Ax$ and let ρ be the smallest of the numbers $|y_i|$ where the y_i are the non-zero scalar components of y. For a sufficiently small ε, $1 > \varepsilon > 0$, let W be defined by

$$|e - x| < \frac{\varepsilon\rho}{\Lambda} \tag{6.7}$$

and if $r = Ae$ then

$$|r-y| < \varepsilon\rho \leq \varepsilon|y| \ . \tag{6.8}$$

Let

$$\delta = \min \left(\frac{2\rho\lambda}{2(2-\gamma)\Lambda}, \alpha\right) \leq \frac{\varepsilon\rho}{2} \tag{6.9}$$

where

$$\alpha = (|r|^2 + \beta)^{1/2} - |r| \tag{6.10}$$

$$\beta = \frac{1}{4}\rho^2\gamma^2(2-\gamma^2)$$

and define the subneighborhood U of x by

$$|e-x| < \frac{\delta}{\Lambda} \ . \tag{6.11}$$

Let $e^{(p)} \in U$ and $e^{(p+1)} \notin W$ then we show that $y_g \neq 0$ for $g = g(p)$. For, if not, then from (6.11) it follows that $|r^{(p)}-y| < \delta$ and $|r_g^{(p)}| < \delta$. From (2.6) and (6.9) it follows that

$$|e^{(p+1)} - e^{(p)}| = \omega_p |B_p r_g^{(p)}| \leq \frac{(2-\gamma)\delta}{\lambda} < \frac{\varepsilon\rho}{2\Lambda}$$

which, together with (6.11), implies that

$$|e^{(p+1)}-x| < \frac{\varepsilon\rho}{\Lambda}$$

This however contradicts the above hypothesis that $e^{(p+1)}$ is not in W. Thus we have that

$$\frac{\rho}{2} < |y_g| - \delta < |r_g^{(p)}| \,. \tag{6.12}$$

Let Q_p be defined by (3.18) and let

$$e(t) = e^{(p)} - tQ_p r^{(p)} , \qquad 0 < t < 1$$

then we show that there exist values of t such that

$$V(x) > V(e(t)) > V(e^{(p+1)}) \,. \tag{6.13}$$

This will contradict the fact that the $V(e^{(p)})$ decrease monotonically to $V(x)$.

We note first that

$$\Delta V \equiv V(e(t)) - V(e^{(p)}) = t\,\omega_p(t\,\omega_p - 2)(r_g^{(p)})^T B_p r_g^{(p)} = k_p \Delta_p V \tag{6.14}$$

and distinguish two cases:

(1) $\qquad \gamma \leq \omega_p < 1$

(2) $\qquad 1 \leq \omega_p \leq 2 - \gamma .$

In case (1) is follows readily that

$$0 < k_p = \frac{t(2-t\omega_p)}{2-\omega_p} < 1$$

or that $\Delta_p V < \Delta V < 0$ and

(6.15) $\qquad V(e^{(p+1)}) < V(e(t)) < V(e^{(p)}) .$

In case (2) if $0 < t < \gamma/(2-\gamma)$ then again $k_p < 1$ and (6.15) holds.

We show now that in both cases there exist values of t such that (6.13) holds. We note first from (6.9) and (6.11) that

(6.16) $$\begin{aligned} V(e^{(p)}) - V(x) &= V(e^{(p)} - x) + 2(e^{(p)} - x)^T y \\ &\leq \Lambda |e^{(p)} - x|^2 + \frac{\delta|r|}{\Lambda} \\ &< \frac{1}{\Lambda} (a^2 + 2a|r|) = \frac{\beta}{\Lambda} \end{aligned}$$

Now, in case (1), if $\gamma < t < 1$ then from (6.12) and (6.14)

(6.17) $$-\Delta V \geq \omega_p t (2-\omega_p t) \frac{\rho^2}{4\Lambda} > \frac{\beta}{\Lambda}$$

which together with (6.16) implies that $V(x) > V(e(t))$.

In case (2), if $\gamma^2 < t < \gamma/(2-\gamma)$, we find, since $\omega_p t < 1$, that $\omega_p t(2-\omega_p t) \geq t(2-t)$ and therefore that (6.13) again is valid. This gives the contradiction and the theorem is proved.

7. Convergence Rate

Let the rate of convergence R be defined, for a relaxation process (2.6) with a fixed partition $\pi = (g_1, \ldots, g_N)$, by

$$(7.1) \qquad R = \inf_{e^{(o)} \in E_n} \left\{ \lim_{t \to \infty} \frac{1}{t} \log \left(\frac{|e^{(o)}|}{|e^{(Nt)}|} \right) \right\}$$

where $t = 0, 1, 2, \ldots$. (This definition is similar to that given in [13]). If this is a r.o.p. of Theorem 1 we then have from (3.22) and (3.24) that

$$(7.2) \qquad R \geq \lim_{t \to \infty} \left(-\frac{1}{t} \log \prod_{p=0}^{Nt} \alpha_p \right) \geq \frac{\lambda}{\mu} \lim_{t \to \infty} \left(\frac{1}{t} \sum_{p=0}^{Nt} (1-\delta_p)\tau_p \right).$$

Since $0 \leq \delta_p < 1$, $0 < \tau_p \leq 1$ the right hand side will exist. If $\tau_p \equiv 1$ and $\delta_p \equiv \delta < 1$ then

$$(7.3) \qquad R \geq \frac{\lambda}{\mu} N(1-\delta) .$$

It would appear from this that δ should be chosen to be zero but this is true (as Ostrowski points out in [6]) only tactically but not for the over-all strategy.

If A is symmetric and positive definite and $g(p)$ is a given ordering, it follows from the proof of (3.20) that

$$V(e^{(o)}) - V(e^{(k)}) = \sum_{p=0}^{k-1} (r_g^{(p)})^T M_p r_g^{(p)} \equiv \Delta_k$$

where

$$M_p = B_p - C_p^T A_{gg} C_p , \qquad g = g(p) .$$

If the C_p satisfy (3.1) and (3.2) then $\Delta_k \geq 0$, $k > 0$. The problem of optimizing the process for a given $r^{(o)}$ and fixed k is therefore equivalent to making a choice for $g(p)$, C_p such that Δ_k = maximum. This "assignment" type of problem seems quite formidable and will not be treated here.

It should be noted that the lower bound given by (7.3) does not give any better result then is obtained, say, in the cyclic process for the discrete Laplacian. This should be expected, however, since a r.o.p. may, at times, be also simultaneously cyclically ordered, as will be seen by examples given below.

8. Examples

We consider the special case of matrices A of order νN of the form

$$A_N = \begin{bmatrix} J & -I & 0 & \cdots & & 0 \\ -I & J & -I & 0 & & 0 \\ 0 & & & & & \\ & -I & J & -I & 0 & \\ & & & & & 0 \\ \vdots & & & & & -I \\ 0 & \cdots & 0 & & -I & J \end{bmatrix}$$

where J is a square matrix of order ν, and I is the unit matrix of order ν, that is, we have the fixed partition case $\pi = (g_1, g_2, \ldots, g_N)$ as described in Remark 2. This form is attained, for example, in the discretization of a Dirichlet problem over a rectangle. The "groups" determined by g_k will correspond to the points on the k^{th} horizontal line.

In the examples to follow we assume that $C_p \equiv 0$ and the norm to be used will be the inner product (4.1). For a given approximation u, N numbers $|r_i|$ are then associated with the N groups. If $|r_m|$ is maximal, then the next approximation is obtained from:

$$\begin{aligned} u'_m &= u_m - Hr_m \\ u'_k &= u_k \qquad\qquad k \neq m \end{aligned} \tag{8.1}$$

where $H = J^{-1}$. If two groups have the same associated norm then choose that one with smallest subscript. (It may also be useful to choose the one nearest the "boundary," that is nearest 1 or N).

The new residuals are then given by:

$$
\begin{aligned}
r_k' &= r_k + Hr_m, && k = m \pm 1,\ 1 < m < N \\
r_k' &= 0 && k = m \\
r_k' &= r_k && k \neq m,\ m \pm 1
\end{aligned}
\tag{8.2}
$$

If $m = 1$ or N the first relation of (8.2) is valid for $k = 2$ or $N - 1$ respectively.

We should note here that the vector Hr_m determines both u' and r' from (8.1) and (8.2).

We will now consider the cases of $N = 2, 3,$ and 5 to estimate $V(e^{(p)})$ and show that the r.o.p. can in each of these cases by cyclic. Before proceeding we note several properties of the matrix A_N.

Let $F_k(\xi)$ be the matrix of order k of the form

$$
F_k(\xi) = \begin{bmatrix}
\xi & -1 & 0 & \cdots & 0 \\
-1 & \xi & -1 & & \\
0 & & & & 0 \\
 & & & & -1 \\
0 & 0 & & -1 & \xi
\end{bmatrix}
\tag{8.3}
$$

and let $\zeta_k(\xi) = \det F_k(\xi)$. Assume that A_N is non-singular then $\zeta_N(J)$ is also non-singular. This follows from the fact that the eigenvalues of $\zeta_N(J)$ are $\zeta_N(\lambda_i)$, $1 \le i \le \nu$, where λ_i is an eigenvalue of J. If for some λ_i, say λ_1, such that $Jy = \lambda_1 y$, $y \neq 0$ we would have $\zeta_N(\lambda_1) = 0$ then there exists a vector $x = (\xi_1,\dots,\xi_N)^T \neq 0$ such that $F_N x = 0$. Let $w = (\xi_1 y^T,\dots,\xi_N y^T)^T$ then $w \neq 0$ and by a simple computation $A_N w = 0$. Since A_N is non-singular this is impossible.

If we now let $A_N^{-1} = G_N = (G_{ij}^{(N)})$, $1 \le i,j \le N$ then a verification, using the identities

$$-\zeta_{k+1}(J) + J\zeta_k(J) - \zeta_{k-1}(J) = 0, \qquad k \ge 1 \tag{8.4}$$

shows that $G_{ij}^{(N)}$ is given by

$$\begin{aligned} G_{ij}^{(N)} &= \zeta_{i-1}(J)\,\zeta_{N-j}(J)\,\zeta_N(J)^{-1} \qquad i \le j \le N+1-i \\ G_{ij}^{(N)} &= G_{ji}^{(N)} = G_{N+1-i,N+1-j}^{(N)} \end{aligned} \tag{8.5}$$

where we set $\zeta_o(J) \equiv I$. The identities (8.4) follow readily from the computation of the $\det F_{k+1}(\xi)$.

If A_N is symmetric and positive definite then $\zeta_k(J)$, $1 \le k \le N$ are non-singular. Let $\psi_k(\xi) = \zeta_k(\xi)/\zeta_{k-1}(\xi)$ then $\psi_k(J)$ are non-singular and we may write $\psi_1(J) = J$, $\psi_{k+1}(J) = J - \psi_k(J)^{-1}$, $k \ge 1$, $\zeta_k(\xi) = \prod_{t=0}^{k} \psi_t(\xi)$ and

$$(8.6)\qquad G_{ij}^{(N)} = \prod_{k=1}^{i-1} \psi_k(J)\,\Big(\prod_{t=N-j+1}^{N} \psi_t(J)\Big)^{-1}, \quad i \le j \le N+1-i.$$

Assume now that $N = 2$ and let a r.o.p. be defined by (8.1) and (4.7). We may clearly assume that the initial residual vector $r^{(o)}$ has the form $r^{(o)} = (x^T,0)^T$ so that the successive residuals are, from (8.2), given schematically by:

i \ p	0	1	2	...
1	x	0	H^2x	
2	0	Hx	0	

or that $r_{m(p)}^{(p)} = H^{(p)}x$. This process is clearly both cyclic and residually ordered.

Since $G_{11}^{(2)} = [\psi_2(J)]^{-1} = (J-H)^{-1}$ and since A_2 is positive definite, the smallest eigenvalue of J is $\lambda(J)=\delta>1$. It follows also from (8.6) that

$$(8.7)\qquad |e^{(p)}| = |A_2^{-1}r^{(p)}| = |H^{p+1}(I-H^2)^{-1}x|$$

and that

$$(8.8)\qquad |e^{(p)}| \le \frac{|x|}{(\sigma^2-1)\sigma^{p-1}} \le \frac{\lambda(A_2)}{(\sigma^2-1)\sigma^{p-1}}\,|e^{(o)}|, \quad p > 0 .$$

Similarly we obtain for the case of $N = 3$ that the resulting r.o.p. is also cyclic if the initial residual has the form $r^{(o)} = (0,x^T,0)^T$. Since $G_{22}^{(3)} = J(J^2-2I)^{-1}$ is

positive definite $\sigma > \sqrt{2}$ and the successive residuals are given by the table:

i \ p	0	1	2	3	. . .
1	0	Hx	0	0	
2	x	0	H^2x	$2H^2x$	
3	0	Hx	Hx	0	

where the order of relaxation is 2,1,3,2,1,3,... since $|H^2x| \leq \frac{1}{\sqrt{2}} |Hx| < |Hx|$. The special form required for the initial residual is no restriction since it can easily be attained by two preliminary relaxations of the first and third group. It follows also from the table and (8.6) that for $p = 3t$, $t = 0,1,2,...$

$$|e^{(p)}| = |A_3^{-1} r^{(p)}| = |H(I-2H^2)^{-1} r_2^{(p)}| = |2^t (I-2H^2)^{-1} H^{2t+1} x|$$

or that

$$|e^{(3t)}| \leq \frac{2^t \Lambda(A_3)}{(\sigma^2-2)\sigma^{2t-1}} |e^{(o)}| , \quad t > 0 . \tag{8.9}$$

For the case $N = 5$ we obtain a r.o.p. which is also cyclic if we assume that the initial residual has the form $r^{(o)} = (0,x^T,0,x^T,0)^T$ and if furthermore $\sigma > 3$. In this case the table obtained is:

i \ p	0	1	2	3	4	5	...
1	0	Hx	Hx	Hx	0	0	
2	x	0	0	$2H^2x$	$3H^2x$	$3H^2x$	
3	0	Hx	2Hx	0	0	0	
4	x	x	0	$2H^2x$	$2H^2x$	$3H^2x$	
5	0	0	Hx	Hx	Hx	0	

where the order used is: $2,4,3,1,5,2,\ldots$. Other possible forms for $r^{(o)}$ may be read off from this table. If $r^{(o)}$ has the symmetry property $r_1^{(o)} = r_5^{(o)}$, $r_2^{(o)} = r_4^{(o)}$ then by three preliminary relaxations of groups $1,5,3$ in that order, the desired form for $r^{(o)}$ is obtained. The error growth can again be estimated for $p = 5t$, $t = 0,1,2,\ldots$ from

$$|e^{(p)}| = |A_5^{-1} r^{(p)}| < \theta\left(\frac{3}{\sigma^2}\right)^t |e^{(o)}| \quad , \tag{8.10}$$

for some positive constant θ depending only on σ .

For these three special cases we obtain from (7.1) that

$$\begin{aligned} N &= 2, \quad R \geq \log \sigma^2 \\ N &= 3, \quad R \geq \log \frac{\sigma^2}{2} \\ N &= 5, \quad R \geq \log \frac{\sigma^2}{3} \quad . \end{aligned} \tag{8.11}$$

For the particular case of the discrete Laplacian, that is, when $J = F_\nu(4)$ it is known that $\sigma = 2 + 4 \sin^2 \frac{\pi}{2(\nu+1)}$, and the rates for $N = 2$ and $N = 3$ are $R \geq \log 4$ and

$R \geq \log 2$ respectively. The rates of convergence of the single step, cyclic, successive relaxation process (also known as Liebmann's method) for these cases are, asymptotically in ν, $\frac{\pi^2}{18}$ and $\frac{\pi^2}{32}$ respectively. Thus we see that the line methods for these cases are better than twice as fast as the Liebmann scheme. This was shown, by Keller[5], to be true for the more general cyclic process using N lines, by another method.

It is generally believed that the scalar residually-ordered processes are not as suitable for automatic digital computers as the cyclic processes. The group r.o.p., on the other hand, being somewhat of a compromise, is more amenable to such use. This method was in fact used in a code, written for the NYU UNIVAC I by G. Rabinowitz, to solve the Dirichlet problem for a rectangle using horizontal lines as the groups. Running time estimates indicate that the r.o.p. converged at better than twice the speed of the cyclic line relaxations, the sacrifice being that both the vectors u and r must be stored simultaneously. There is only a small increase in computation since, as pointed out above, both u' and r' are obtainable from the one vector Hr_m. The r.o.p. may also be more useful where greater accuracy is needed, since error estimates are available for any finite number of relaxations made.

Bibliography

[1] G. Birkhoff and R. S. Varga, Reactor Criticality and non-negative matrices, WAPD-166 (July 1957).

[2] L. Collatz, Einige Anwendungen Funktionanalytischer Methoden in der Praktischen Analysis, Z. Angew. Math. Physik vol 4. (1953), pp. 327-357.

[3] G. E. Forsythe, What are relaxation methods? "Modern Mathematics for the Engineer," E. F. Beckenbach, Ed., McGraw-Hill, New York 1956, pp. 428-447.

[4] C. F. Gauss, Brief an Gerling, Dec. 26, 1823, Werke, Vol. 9 pp. 278-281. (Translated by G. E. Forsythe, MTAC, Vol. 5, (1951), pp. 255-258).

[5] H. B. Keller, On some iterative methods for solving elliptic difference equations, AEC Computing Center, N.Y.U., Report NYO-7971 (1957).

[6] A. M. Ostrowski, On the linear iteration procedures for symmetric matrices, Rendic. di Mat. e.d.s. Applicaz. (5) vol. 13 (1954), pp. 140-163.

[7] _____, Determinanten mit Uberwiegender Hauptdiagonale und die Absolute Konvergenz von Linearen Iterationsprozessen, Comment. Math. Helv. vol. 30 (1956), pp. 175-210

[8] G. de Rham, Sur un Théorème de Stieltjes Relatif à Certaines Matrices, Acad. Serbe Sci. Publ. Inst. Math. vol. 4 (1952), pp. 133-134.

[9] R. V. Southwell, Relaxation Methods in Theoretical Physics, Clarenden Press, Oxford, 1946.

[10] T. J. Stieltjes, Sur les racines de l'équation $X_m = 0$, Acta Math. vol. 9 (1886), pp. 385-400.

[11] G. Temple, The general theory of relaxation methods applied to linear systems, Proc. Roy. Soc. London, Ser. A, vol. 169 (1939), pp. 476-500.

[12] J. Todd, The condition of certain matrices, III, J. Research Nat. Bur. Standards vol. 60 (1958), pp. 1- 7.

[13] W. Kahan, The rate of convergence of the extropolated Gauss - Seidel iterations (abstract) J. Assn. Comp. Mach. vol. 4 (1957), pp. 521-522.

[14] H. Geiringer, On the solution of linear equations by certain iteration methods, Reissner Anniversary Volume, J. E. Edwards, Ann Arbor, 1949, pp. 365-393.

[15] D. Young, Iterative methods for solving partial difference equations of elliptic type, Trans. Amer. Math. Soc. vol. 76 (1954) pp. 92-111.

[16] R. J. Arms, L. D. Gates, and B. Zondek, A method of block iteration, J. Soc. Indust. Appl. Math. vol. 4, (1956), pp. 220-229.

[17] S. R. Frankel, Convergence rates of iterative treatments of partial differential equations, MTAC Vol. 4 (1950) pp. 65-75.

8673 Schechter
Relaxation methods for

NYU
NYO
8673
c.1

NYU
NYO
8673
c.1

AUTHOR Schechter

TITLE Relaxation methods for li-
near equations.

DATE DUE | BORROWER'S NAME | ROOM NUMBER

N. Y. U. Institute of
Mathematical Sciences
25 Waverly Place
New York 3, N. Y.

JAN 2 2 1959

Date Due

PRINTED IN U. S. A.